AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration

Year: 2023 | Citations: 908 | Authors: Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang

Abstract

Large language models (LLMs) have transformed numerous AI applications. On-device LLM is becoming increasingly important: running LLMs locally on edge devices can reduce cloud computing costs and protect users' privacy. However, the astronomical model size and the limited hardware resources pose significant deployment challenges. To solve these issues, we propose Activation-aware Weight Quantization (AWQ) and TinyChat, an algorithm-system full-stack solution for efficient on-device LLM deployment. AWQ is a novel quantization method that identifies and protects salient weights based on activation distribution, significantly reducing model size while preserving performance. TinyChat, an optimized inference framework, translates AWQ's theoretical memory savings into practical speedups through techniques such as on-the-fly dequantization, SIMD-aware weight packing, and kernel fusion. Together, they enable 4x model size reduction and 3-4x acceleration across various edge platforms, from high-end desktop GPUs to resource-constrained IoT devices. This solution democratizes on-device LLM deployment, offering privacy-preserving, low-latency AI capabilities across a wide range of applications.